Journal of Applied Psychology
1969, Vol. 53, No. 6, 439-445

PREDICTION OF JOB PERFORMANCE FROM ASSESSMENT REPORTS:

USE OF A MODIFIED *Q*-SORT TECHNIQUE TO EXPAND PREDICTOR AND CRITERION VARIANCE [1]

GARLAND Y. DeNELSKY [2] AND MICHAEL G. McKEE [3]

Central Intelligence Agency

Predictions of performance and personality characteristics made on the basis of preemployment psychological assessment reports were compared with subsequent performance evaluations contained in the fitness reports of 32 government employees. Seven psychologists reviewed the assessment reports as a basis for predicting overall job effectiveness and specific performance and personality characteristics. They then reviewed the narrative section of each individual's fitness report as a basis for rating the overall effectiveness of each person. Ratings were made using a modified *Q*-sort technique that reliably expanded the variances of predictor and criterion variables. A significant positive relationship was found between predicted and actual effectiveness. In addition, the psychologists were able to predict specific performance and personality dimensions on a significantly better than base-rate basis.

Over the past 20 years, with the 1948 Office of Strategic Services volume, *Assessment of Men,* lighting the way, there has been a steady if slow flow of research on the predictive validity of clinical assessment, using multiple methods for obtaining information about individuals. Taft (1959) provides a comprehensive review of the earlier studies. Studies by Bray and Grant (1966), Hilton, Bolin, Parker, Taylor, and Walker (1955), Campbell, Otis, Liske, and Prien (1962), Trankell (1959), Dicken and Black (1965), and Albrecht, Glaser, and Marks (1964) report significant positive correlations between assessment predictions and performance criteria. The results of some studies, however, have cast doubt upon the predictive efficacy of assessment procedures (Holtzman & Sells, 1954; Kelly & Fiske, 1951).

Bray and Grant (1966) summarized the research to date as follows:

> Though no firm conclusions regarding the predictive validities of multiple assessment procedures can be drawn from the rather mixed findings of published research, it does appear clear that the more accurate predictions were obtained where the performance to be predicted was clearly defined, the assessment results did not restrict the range of subsequent criterion performance, and the criterion measures employed were not limited by low reliability and questionable validity [p. 2].

Unfortunately, it is usually impossible to meet the above conditions in applied assessment; the job duties are heterogeneous and ill defined; criterion performance is restricted in range by selection on the basis of assessment results; the criterion measure is based on standard organizational evaluation reports and, as such, is of questionable validity. A variety of raters and a variety of jobs, with

[1] The views expressed in this article are those of the authors and do not necessarily reflect an official position of the Central Intelligence Agency.

[2] Requests for reprints should be sent to Garland Y. DeNelsky, Central Intelligence Agency, Washington, D. C. 20505.

[3] Now at the Cleveland Clinic, Cleveland, Ohio.

the clearly inept performers screened out, tend to lower the correlations of predictors and rated job performance. Many elements in a study of assessment *au naturel* coalesce to lower validity, and the question is whether assessment has value within these limitations and whether it can predict performance in an ongoing occupational setting.

The purpose of the present study was to determine if predictive validity can be demonstrated for psychological assessments within a natural setting when a special rating technique that increases predictor and criterion variability is used. The specific focus of investigation was the assessment report; the major question was whether preemployment psychological assessment reports do predict the subsequent performance of those individuals who are hired.

Method

Subjects

Fitness reports (routine performance evaluations about one-half page in length) were obtained on 32 male employees who had been working overseas for 1 yr. or more. Assessment reports were available on all 32. These individuals had been assessed 12–57 mo. earlier by one of eight psychologists; the median interval between assessment and fitness reports was 20 mo. The original assessments varied slightly from case to case but typically included intellectual, personality, attitudinal, and interest testing in addition to one or more depth interviews. The assessment reports were typically one or two pages long and contained descriptions of the individual's strengths and weaknesses as well as a summary recommendation.

All 32 men were overseas at the time their fitness reports were prepared. Although it was not possible to determine how many different supervisors had actually been responsible for this group, it was established that none of the field supervisors had seen their assessment reports. The total of 32 men was divided into two groups. Each of these groups (which will be referred to as Group 1 and Group 2) contained 16 men. The two groups were judged separately; in fact, several months intervened between the judging of Group 1 and Group 2.

Seven staff psychologists served as judges. All had experience in assessing overseas candidates.

Procedure

Trait prediction. In the first phase of the study for both groups, each of the judges was given the 16 original assessment reports, together with a specially designed Trait Rating Sheet for each *S*. The Trait Rating Sheet listed 25 performance and personality traits that had been abstracted from the narrative sections of the total group of fitness reports of the employees in the study. Performance ratings included such dimensions as response to supervision, accuracy of work, speed of learning, and supervisory effectiveness; personality ratings included such dimensions as judgment, maturity, flexibility, and self-confidence. Approximately half of the 25 dimensions could be described as personality variables; the other half pertained to job performance. The judges were instructed to form an impression of each of the men from the assessment report, and, on the basis of this impression, to predict whether each individual would be discussed favorably or unfavorably on each trait in his fitness report (assuming, of course, that he would be discussed on all dimensions—a slightly unrealistic situation since no employee was mentioned on more than 12 of the 25 dimensions). For those individuals mentioned favorably or unfavorably on a given dimension in their fitness reports, it was possible to determine if the predictions made by psychologists were in the same direction as the actual descriptions of the individuals in their fitness reports.

Q *sorts of assessment and fitness reports.* Following his completion of the Trait Rating Scales, each judge was asked to sort the assessment reports of the 16 men of each group into five categories corresponding to his prediction of each individual's overall effectiveness in a typical overseas work situation of the type to which these men were assigned. In order to eliminate variance due to differing frames of reference on the part of the seven judges, a modified *Q*-sort distribution was used; assessment reports were to be assigned to five categories, ranging from a predicted worst performance to a predicted best performance with 1, 4, 6, 4, and 1 individuals assigned to the respective categories. Score values of 1, 2, 3, 4, and 5 (best) were assigned to the five categories.

Following the *Q* sort of assessment reports on the basis of predicted overall effectiveness, each judge was assigned the task of *Q* sorting, in the same manner as before, each group of 16 individuals on the basis of actual overall effectiveness as described in narrative form in their fitness reports. The names of the 16 men were deleted from the fitness reports; thus the judges had no way of knowing which of the assessment reports and fitness reports had been written for the same persons.

It should be noted that the prediction situation as structured in this study was different from the usual design of studies with similar objectives. Instead of being given test scores and other psychometric and background data and being required to weight this "raw" information in order to make predictions of future behavior, the judges in this study were asked to formulate predictions on the basis of finished assessment reports. Thus, the judges in the present study were placed in a role similar to the consumer of psychological assessment reports: They were to make predictions on the basis of someone else's analysis and interpretation of first-hand data. Dicken and Black (1965) used a similar method,

TABLE 1

ANALYSIS OF VARIANCE RELIABILITY COEFFICIENTS FOR ASSESSMENT- AND FITNESS-REPORT RATINGS

Rating	Coefficient for single rating		Coefficient for composite rating	
	Group 1	Group 2	Group 1	Group 2
Assessment report	.63	.66	.92	.93
Fitness report	.59	.74	.91	.95

commenting that "the ratings are thus two interpretive steps removed from the original test data [p. 36]."

RESULTS

Prediction of Overall Effectiveness

Before relating assessment-report predictions to fitness-report ratings, it was necessary to establish the reliability of the judgments made by the judges on both measures.

Table 1 presents the analysis of variance reliability coefficients for the assessment- and fitness-report judgments. It is evident from this table that the reliabilities, particularly of the average or composite ratings for each individual by all judges, are quite satisfactory. Despite several judges' comments that the task of making the ratings was a difficult one, there was substantial agreement among judges on both the assessment-report and the fitness-report ratings.

The answer to the primary question of this study—whether judges can predict, on the basis of psychological assessment reports, performance in actual field situations as judged from fitness-report narratives 12–57 mo. later—can be approached from a number of directions. Perhaps the single most meaningful approach is to correlate the composite assessment-report predictions of the seven judges for each of the 16 individuals in each group with the composite judged effectiveness of the same individuals based on fitness reports. The resulting correlations, presented in Table 2, indicate that with the total sample of 32 men, there is a significant positive relationship between the overall or composite predictions of effectiveness based on assessment reports and actual effectiveness as judged from fitness reports.

TABLE 2

CORRELATIONS BETWEEN COMPOSITE ASSESSMENT-REPORT PREDICTIONS AND FITNESS-REPORT EVALUATIONS

Group	N	r
1	16	.42
2	16	.25
1 and 2 combined	32	.32*

* $p < .05$, one-tailed test.

Another way of illustrating the relationship between assessment and fitness reports is shown in Table 3. Of those 17 men with average or above assessment ratings, 12 (71%) received average or above fitness ratings, while only 6 (40%) of the 15 men with below-average assessment ratings received average or above fitness ratings.

Table 4 presents correlations between the individual judge's assessment ratings and the composite fitness ratings (for Groups 1 and 2 combined). Assuming the composite of the fitness-report ratings by all judges is the best single measure of actual performance, the psychologists varied in their ability to predict performance from assessment reports; only three of the correlations were significant at the .05 level.

The fitness reports used in this study required the evaluator not only to give a narrative appraisal but to rate the overall performance of each of his subordinates on a 5-step adjectival scale: weak, adequate, strong, proficient, outstanding. In this study, the adjectival ratings were not made available to the judges since it was thought that differences in rating might reflect variations in

TABLE 3

PERFORMANCE AS A FUNCTION OF ASSESSMENT PREDICTION

Assessment prediction	Performance evaluation	
	Average or above	Below average
Average or above[a]	71%	29%
Below average[b]	40%	60%

[a] $N = 17$.
[b] $N = 15$.

TABLE 4

Correlations between Individual Assessment-Report Predictions and Composite Fitness-Report Evaluations

Judge	Correlations between individual ratings of assessment reports & composite (7 judges) fitness-report ratings
1	.29
2	.30*
3	.30*
4	.19
5	.41*
6	.22
7	.13

* $p < .05$, one-tailed test.

rating bias of raters more than variations in performance. Table 5 presents data indicating that the judges in this study evaluated the narrative section of the ratee's fitness reports in the same direction as the overall letter ratings assigned to each man by his supervisor. Remembering that the larger the numerical rating an individual received the higher was his judged effectiveness, individuals receiving overall "strong" ratings were judged more effective than those receiving overall "proficient" ratings ($p < .07$). The biserial correlation between the judged composite rating of effectiveness and the overall letter rating was .34. More important than the agreement of supervisors' ratings of overall performance with the judges' ratings based on the supervisor's narrative evaluation is the fact that the judges' ratings provide a greater range than is usually obtained with fitness reports in which the majority of supervisors generally restrict themselves to about two categories, as they did in this study where all the overall ratings were either proficient or strong. The high reliability of the 5-point ratings made by the psychologists suggests that a greater range of performance among personnel is recognized by supervisors than is typically reflected in their overall ratings in fitness reports.

TABLE 5

Mean Effectiveness Ratings for Individuals Receiving Strong and Proficient Overall Fitness-Report Evaluations

Item	Mean composite effectiveness rating[a]
Individuals receiving overall strong fitness-report evaluations[b]	22.3
Individuals receiving overall proficient fitness-report evaluations[c]	19.1

Note.—An evaluation of "strong" was superior to "proficient" in the fitness-reporting system.
[a] As judged by seven psychologists from fitness-report narratives only.
[b] $N = 19$.
[c] $N = 13$.

Trait Prediction

In this portion of the study, the seven psychologists, on the basis of assessment reports only, rated all 32 employees on 25 traits or dimensions that had been abstracted from the fitness reports of the total group of individuals. Using the specially designed Trait Rating Sheet, judges predicted whether each individual would be discussed favorably or unfavorably on each dimension in his fitness report, assuming that he would be discussed on all dimensions.

A major difficulty with these data arose because 88% of the 188 statements abstracted from the fitness reports of the 16 individuals were favorable. Similarly, 74% of the total number of predictions made by the judges were positive. These high-positive base rates insured a great deal of agreement between predictions based on assessment reports and statements drawn from fitness reports. In fact, 74% of the total group of over 1,300 predictions made by the seven psychologists were "correct," that is, in agreement with the fitness-report narratives. Given the high rate of positive statements in fitness reports and the nearly as high rate of positive predictions made from assessment reports, were the psychologists able to make a significant improvement over the base rates in their prediction of these specific dimensions of performance?

One way of answering this question is presented in Table 6. If psychologists are able to predict specific dimensions of performance to a degree exceeding that which would be expected by base rates alone, then their predictions for those individuals described positively in fitness reports on a specific dimension should exceed the overall (or base rate) prediction for all persons on that dimension. Since for most dimensions the distribution of the psychologists' predictions was skewed, the median rather than the mean percentage of psychologists' predictions of favorable fitness-report descriptions on a given dimension was taken as the base rate for that dimension. For example, if 85% of the judges predicted that a certain individual would be described favorably on a given dimension and in fact he was described favorably in his fitness report on this dimension, this would constitute a successful prediction if the median percentages of judges rating all individuals positively on that dimension was 71. If, however, only 57% of the judges predicted that this person would receive favorable mention on this dimension, this would be classified as an unsuccessful prediction since it is below the 71% base rate. But if this person's fitness report had made an *unfavorable* comment about his initiative and resourcefulness, the first prediction (where 85% of the judges predicted a favorable description) would have been classified as unsuccessful since it was above the base rate while the second prediction would be successful (since only 57% of the judges predicted a favorable description of this dimension as compared with a base rate of 71%). This is a rather rigorous test, for it assumes that people mentioned favorably in their fitness reports on a specific dimension are actually stronger, and the people mentioned unfavorably, weaker on that dimension than people not mentioned one way or the other. The typical fitness report, of course, does not provide a comprehensive or systematic picture of a person's strengths or weaknesses.

Table 6 shows that for 83 of the total group of 150 positive statements drawn from fitness reports, the group of seven psychologists made predictions on the corresponding dimensions that were more in the correct (or favorable) direction than the average of the total group of predictions made on these dimensions. Similarly, for the 21 negative statements drawn from the fitness reports, the psychologists made 16 correct predictions on the corresponding dimensions. Thus, for a combined total of 99 of 171 predictions, the psychologists achieved more accurate predictions than would have been expected through base rates alone. A binomial test indicates that this ratio of successful to unsuccessful predictions exceeds a .50:.50 (chance) split at the .02 level. (Seventeen positive statements drawn from fitness reports could not be classified as successful or unsuccessful predictions since the percentage of psychologists predicting a favorable fitness-report description fell at the median for all *S*s on those dimensions.)

TABLE 6

Number of Successful and Unsuccessful Predictions Made on Specific Performance and Personality Dimensions Described in Fitness Reports

Dimension	Successful predictions	Unsuccessful predictions	Total
Positive	83	67	150
Negative	16	6	21
Total	99*	72	171

Note.—"Successful" and "unsuccessful" were defined in terms of base rates; a successful prediction for an individual on a given dimension was recorded when the percentage of judges rating that individual in the same direction as the fitness report's narrative exceeded the median percentage of the judges rating all individuals on that dimension. (See the text for a complete description of this method.)

* $p < .02$ that this split is significantly different from a .50:.50 split.

Because of the relatively few individuals discussed on each of the various dimensions of the Trait Rating Scale in the fitness reports (no more than 20 of 32 individuals were cited on any single dimension), it is not possible to compare the relative predictive effectiveness of the group of psychologists on different dimensions. However, there is evidence that the psychologists in this study were better able to predict weaknesses than strengths. On positive dimensions, 55% of the psychologists' predictions were successful (i.e., better than the base rates). On negative dimensions, 76% of their predictions were successful. The difference between these proportions was significant at the .05 level.

Discussion

On the basis of this study, it is reasonable to conclude that psychologists can predict significantly better than chance both overall competence and specific performance and personality characteristics of employees using only completed assessment reports prepared 1–4 yr. earlier.

The modest relationships that emerged for the prediction of overall as well as specific dimensions of effectiveness are probably artificially low, since the least promising individuals were not employed at all. This type of restriction of range is unavoidable in studies of this nature. Had it been possible to gather feedback data on all individuals assessed, it is likely that the predictive effectiveness of the psychologists would have been enhanced.

It was found that the pooled judgments of several judges yielded greater predictive accuracy than the judgments of individual psychologists. Only one of the seven judges was able to exceed the predictive accuracy of the composite judgments. As Kelley and Thibaut (1954) point out, pooling independent judgments should always enhance validity except in the situation where the judgments of the average individual correlate zero with the criterion.

The finding that psychologists were able to predict specific performance dimensions and personality characteristics better than the base rate was encouraging. It should be remembered that these predictions were made on the basis of secondary information; that is, the psychologists who made the predictions used assessment reports that were not formulated specifically toward making predictions on these dimensions. Therefore, the psychologists in this study were forced to "read between the lines" to make predictions on most of the dimensions for most of the employees. Higher predictive accuracy could be expected if the psychologists who made the predictions conducted the initial assessments with these dimensions in mind.

The finding that psychologists were better able to predict weaknesses than strengths is provocative. If substantiated by further research, it has interesting implications for the assessment process.

That psychologists can reliably generate 5-point evaluations of fitness reports that originally fell in only two categories is noteworthy. One of the difficulties in using many standard fitness reports or appraisal ratings as criteria of job performance is their limited variance. The results of this study indicate that job-performance variance can be meaningfully expanded through a modified *Q* sort that forces reviewers of these reports to make more discriminations among individuals.

Finally, studies similar to the present one should be conducted with persons other than psychologists making predictions on the basis of assessment reports. This would be more nearly analogous to the situation at present where the psychologist, through his assessment report, supplies a consultative function to another individual (or group of individuals) who combines this report with other information in order to arrive at a selection decision. Implicit in this decision is the prediction of how well a given individual will "work out," or even whether he will "work out" at all. In the last analysis, these predictions made by the persons who typically select or reject are the most meaningful ones, and hence should be the focus of systematic study.

Meanwhile, this study does provide reassurance that the assessment process can result in meaningful predictions of job behavior as evaluated from fitness reports.

REFERENCES

ALBRECHT, P. A., GLASER, E. M., & MARKS, J. Validation of a multiple assessment procedure for managerial personnel. *Journal of Applied Psychology,* 1964, **48**, 351–360.

BRAY, D. W., & GRANT, D. L. The assessment center in the measurement of potential for business management. *Psychological Monographs: General and Applied,* 1966, **80**(**17**, Whole No. 625).

CAMPBELL, J. T., OTIS, J. L., LISKE, R. E., & PRIEN, E. P. Assessment of higher-level personnel: II. Validity of the overall assessment process. *Personnel Psychology*, 1962, **15**, 63–74.

DICKEN, C. F., & BLACK, J. D. Predictive validity of psychometric evaluations of supervisors. *Journal of Applied Psychology*, 1965, **49**, 34–37.

HILTON, A. C., BOLIN, S. F., PARKER, J. W., JR., TAYLOR, E. K., & WALKER, W. B. The validity of personnel assessments by professional psychologists. *Journal of Applied Psychology*, 1955, **39**, 287–293.

HOLTZMAN, W. H., & SELLS, S. B. Prediction of flying success by critical analysis of test protocols. *Journal of Abnormal and Social Psychology,* 1954, **49**, 485–490.

KELLEY, H. H., & THIBAUT, J. W. Experimental studies of group problem solving and process. In G. Lindzey (Ed.), *Handbook of social psychology*.

Vol. 2. *Special fields and applications.* Cambridge: Addison-Wesley, 1954.

KELLY, E. L., & FISKE, D. W. *The prediction of performance in clinical psychology.* Ann Arbor: University of Michigan Press, 1951.

OSS Assessment Staff, *Assessment of Men.* New York: Rinehart, 1948.

TAFT, R. Multiple methods of personality assessment. *Psychological Bulletin,* 1959, **56**, 333–352.

TRANKELL, A. The psychologist as an instrument of prediction. *Journal of Applied Psychology,* 1959, **43**, 170–175.

(Received December 30, 1968)